IThOS

User Guide : version 2.0

Nouri Ben Zakour
Yves Le Loir
Dominique Lavenier

---

1. Introduction

IThOS is a software package dedicated to the design of primers. The input is one or several genome sequences (FASTA files) and the output is a list of primer candidates that fulfill criteria set by the user. IThOS also determines putative hybridization sites for these primers.

IThOS works in a two-step procedure:

1. Primer design
2. Verification of hybridization sites

Each step relies on a dedicated program:

- Step 1 : `ithos_gen <genome> <parameters> <primers> [-c]`
- Step 2 : `ithos_chk <genome> <primers> <parameters> <pr_out> <pr_hyb> [-n]`

Both programs can be used separately. A third program completes the package and enables the visualization of the primer features:

- `ithos_viz <primers> <parameters>`

Parameters of the different programs are:

`<genome>`
: File containing one or several DNA sequences, FASTA format.

`<parameters>`
: File specifying criteria for primer selection.

`<primers>, <pr_in>, <pr_out>, <pr_hyb>`
: Files containing a primer list, FASTA format.

`[-c]`
: Option enabling the search of primers on the complementary strand.

`[-n]`
: Option indicating that ithos_chk is used after ithos_gen: the same genome sequence is given as an input and the output file <primers> of ithos_gen is given as an input to ithos_chk without modification (to make sure that the real hybridization sites for the selected primers are not being considered as secondary hybridization sites!)

2. Design of the primers: ithos_gen

Starting from one or several DNA sequences, the goal is to design primers that fulfill criteria set by the user. The ithos_gen program considers all the words whose size is in an interval that corresponds to the minimum and maximum primer length. For each word, a suite of filters is applied. All oligonucleotides passing successfully through the filters are proposed as primer candidates. For each filter, several parameters can be set by the user to refine the primer selection according to the application. Six filters are implemented. They are described in the following sections.

Filter 1: G+C %

For a primer of size T, filter 1 works as follows:

- Counts the number of G and C nucleotides (# GC)
- Calculates the percentage ➡ P = (# GC * 100) / T
- Discard the primer if: P < pcGCMin **or** P > pcGCMax

The default values are:

- pcGCMin = 40
- pcGCMax = 60

Filter 2: Tm (*Melting temperature*)

The nearest neighbor method is used to calculate the primer (Santa Lucia et al., 1998). It also takes into account the concentration of nucleotides (dnaConc) and the concentration of salt (saltConc).

This filter removes primers if:

- Tm < oligoTmMin
- Tm > oligoTmMax

The default values are:

- oligoTmMin = 57 °C
- oligoTmMax = 62 °C
- dnaConc = 500 nM
- saltConc = 50 nM

Bibliography

A Unified View of Polymer, Dumbbell, and Oligonucleotide DNA Nearest-Neighbor Thermodynamics John SantaLucia, *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 95, No. 4 (Feb. 17, 1998), pp. 1460-1465

Filter 3: Number of repeats

This filter removes oligonucleotides with N consecutive identical nucleotides or dinucleotides (nbRepeat). For example, if nbRepeat = 4, the following primers will be removed:

```
1: GGGATGGACACGGA TTTT GGACCAGC
2: TTAGC TATATATA GGCAGGGATTAGG
```

The first one presents a suit of 4 « T ». The second one a suite of 4 « TA »
The default value is:

- nbRepeat = 5

Filter 4: Hairpin

This filter removes oligonucleotides with hairpin loops that present the following features:

- stem size > or = to maxHpDup
- loop size > or = to MaxHpLoop

loop

stem

The default values are:

- maxHpDup = 4
- maxHpLoop = 4

Filter 5 : self-complementarity

This filter checks that a primer will not hybridize with itself during PCR. Thus, primers that form a duplex with their complementary strand are removed. Criteria for selection are as follow:

- For the full length of the pirmer, the authorized deltaG value must not exceed maxDeltaGAuto
- For the 3' end and a distance of sizeDeltaGAuto3, the authorized deltaG value must not exceed maxDeltaGAuto3.
- The minimal size of a tested duplex is sizeDeltaGAuto

The default values are:

- maxDeltaGAuto = -10kcal/mol
- maxDeltaGAuto3p = -7kcal/mol
- sizeDeltaGAuto = 6
- sizeDeltaGAuto3p = 8

Filtre 6 : thermodynamic stability at primer ends

This filter calculates a deltaG on the 5’ and 3’ ends of the primers. The size of the 5’ end to be considered is given by sizeExt5. The size of the 3’ end to be considered is given by sizeExt3.

A primer is removed if:

- The value in 5’ is above deltaG5
- The value in 3’ is out of the interval [deltaG3Min,deltaG3Max]

The default values are:

- sizeExt5 = 5
- sizeExt3 = 5
- deltaG5 = - 4 kcal/mol
- deltaG3Min = -6 kcal/mol
- deltaG3Max = - 4 kcal/mol

Software execution

The software is launched by the command line:

```
ithos_gen <genome> <parameters> <primers> [-c]
```

input files

- `<genome>` is a text file (FASTA) that contains a genome that can be cut into several sequences.
- `<parameters>` is a text file that contains the parameters for the different filters

output files

- `<primers>` is a text file (FASPTA format) that contains all the selected primers. Coordinates of the selected primers are given as comments in between brackets ([ ]).
- The software generates an additional file `<primers.log>` that gives a few statistical data on the filtering process.

Option

- `[-c]` allows researching on the complementary strand of the genome

Example 1 : search for primers on the leading strand

If a file containing the 2 following sequences is considered:

```
>sequence #1
CGATTAAAGATAGAAATACACGATGCGAGCAATCAAATTTCA

>sequence #2
GAAACAACAAAACCTTCTACTGAAACAACTGAGGATAAT
CATGTGCTTGGTAGAGAGCAATTCAATGCCC
```

The software ithos_gen will produce the following file:

```
>sequence #1    [8 31]
GATAGAAATACACGATGCGAGCAA
>sequence #1    [10 33]
TAGAAATACACGATGCGAGCAATC
>sequence #1    [11 34]
AGAAATACACGATGCGAGCAATCA
>sequence #1    [12 35]
GAAATACACGATGCGAGCAATCAA
>sequence #1    [17 40]
ACACGATGCGAGCAATCAAATTTC
>sequence #1    [18 41]
CACGATGCGAGCAATCAAATTTCA
>sequence #2    [27 50]
ACTGAGGATAATCATGTGCTTGGT
>sequence #2    [44 67]
GCTTGGTAGAGAGCAATTCAATGC
```

The numbers in brackets indicate the coordinates – begin and end – of the primer in the sequence.

Example 2 : search for primers on the complementary strand (option –c)

The search for primers is done on the complementary strand of the genome sequence. The software generates exactly the same type of file. The only difference is that the primer coordinates are inverted. Thus, if we consider the previous example, the program will generate:

```
>sequence #1    [35 12]
TTGATTGCTCGCATCGTGTATTTC
>sequence #1    [34 11]
TGATTGCTCGCATCGTGTATTTCT
>sequence #1    [33 10]
GATTGCTCGCATCGTGTATTTCTA
>sequence #1    [31 8]
TTGCTCGCATCGTGTATTTCTATC
>sequence #1    [30 7]
TGCTCGCATCGTGTATTTCTATCT
>sequence #1    [29 6]
GCTCGCATCGTGTATTTCTATCTT
>sequence #2    [69 46]
GGGCATTGAATTGCTCTCTACCAA
>sequence #2    [68 45]
GGCATTGAATTGCTCTCTACCAAG
>sequence #2    [67 44]
GCATTGAATTGCTCTCTACCAAGC
>sequence #2    [66 43]
CATTGAATTGCTCTCTACCAAGCA
>sequence #2    [50 27]
ACCAAGCACATGATTATCCTCAGT
```

3. Checking the hybridization sites: ithos_chk

For each primer, the program checks that there are no secondary hybridization sites elsewhere on the whole genome, e.g. for a primer: `5′ - T G A – 3′`, the following hybridization sites must be checked:

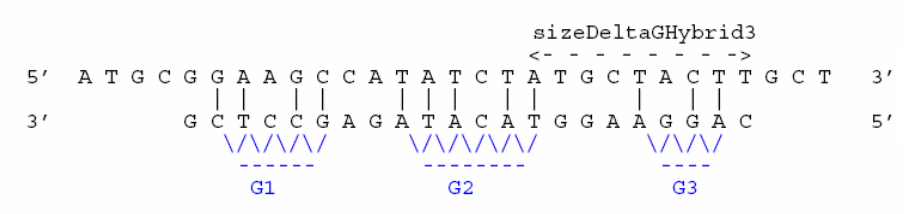
genome

Detection of a secondary hybridization site is not based on the percentage of identity but on the calculation of the thermodynamic stability of the duplex (cf. filter auto-complementarity).

A maximal deltaG is calculated on the whole primer length and a maximal deltaG in 3' is calculated on sizeDeltaGHybrid3 nucleotides. A deltaG value is calculated on consecutive matches, including putative mismatches (1 mismatch surrounded by 2 matchs). A hybridization site is recognized if one of the two conditions is true:

- deltaG < maxDeltaGHybrid
- deltaG in 3' < maxDeltaGHybrid3

For example:



primer
genome

deltaG = max (G1,G2,G3) and deltaG in 3' = G3

The values for G1, G2 and G3 are the sum of the thermodynamic values between 2 consecutive nucleotide pairs.

The default values are:

- maxDeltaGHybrid = -16 kcal/mol
- maxDeltaGHybrid3 = - 9 kcal/mol
- sizeDeltaGHybrid3 = 8

Software execution

The software is launched by the following command line:

```
ithos_chk <genome> <primers> <parameters> <pr_out> <pr_hyb> [-n]
```

Input files

- `<genome>` is a text file (FASTA format) containing a genome sequence

• <primers> is a text file (FASTA format) containing a list of primers
• <parameters> is a text file containing the parameters for the different filters

Output file

• <pr_out> a text file containing all the primers that have no hybridization sites
• <pr_hyb> a text file containing all the primers that have at least one hybridization site
• The program generate an additional file <pr_hyb.info> that indicates, for each primer of the file <pr_hyb> the positions of hybridization on the genome as well as the deltaG values.

Option

• [-n] indicates that ithos_chk is used after ithos_gen: the same genome is given as input file and the output file <primers> of ithos_gen is given as input for ithos_chk without modification (this way, the own coordinates of the primers are not considered as secondary hybridization sites!)

Example 3: Primer design on a genome and elimination of the primers that hybridize at other positions

If the following genome is considered and memorized in a file named ex3

```
>exemple_3
AAGATAGAAATACACGATGCGAGCAATCAAATTTCAGGTAGAAAGGATAGA
AATACACGAAGCGAGCAATCAACTTGACCTAGGTGAGGGATAGGACCAGA
```

Primer design is launched by the command line:
`ithos_gen ex3 parameters primer`

This gives a file named primer. This file contains, for example, the following primer list:

```
>exemple_3 [5 29]
AGAAATACACGATGCGAGCAATCAA
>exemple_3 [11 35]
ACACGATGCGAGCAATCAAATTTCA
>exemple_3 [13 37]
ACGATGCGAGCAATCAAATTTCAGG
>exemple_3 [44 68]
GGATAGAAATACACGAAGCGAGCAA
>exemple_3 [49 73]
GAAATACACGAAGCGAGCAATCAAC
>exemple_3 [59 83]
AAGCGAGCAATCAACTTGACCTAGG
>exemple_3 [64 88]
AGCAATCAACTTGACCTAGGTGAGG
>exemple_3 [70 94]
ACTTGACCTAGGTGAGGGATAGGAC
>exemple_3 [76 100]
GACCTAGGTGAGGGATAGGACCAGA
```

Checking of the hybridization sites is carried out by the command line:
`ithos_chk ex3 primer parameters primer_ok primer_hyb -n`

This produces 3 output files:

primer_ok

```
>exemple_3 [59 83]
AAGCGAGCAATCAACTTGACCTAGG
>exemple_3 [64 88]
AGCAATCAACTTGACCTAGGTGAGG
>exemple_3 [76 100]
GACCTAGGTGAGGGATAGGACCAGA
```

primer_hyb

```
>exemple_3 [5 29]
AGAAATACACGATGCGAGCAATCAA
>exemple_3 [11 35]
ACACGATGCGAGCAATCAAATTTCA
>exemple_3 [13 37]
ACGATGCGAGCAATCAAATTTCAGG
>exemple_3 [44 68]
GGATAGAAATACACGAAGCGAGCAA
>exemple_3 [49 73]
GAAATACACGAAGCGAGCAATCAAC
>exemple_3 [70 94]
ACTTGACCTAGGTGAGGGATAGGAC
```

Primer_hyb.info

```
>exemple_3 [5 29]
AGAAATACACGATGCGAGCAATCAA
                                  seq   start  end    dG max       dG 3'
5' AGAAATACACGATGCGAGCAATCAA 3' primer
|||||||||||| ||||||||||||
3' TCTTTATGTGCTTCGCTCGTTAGTT 5' genome 48     72     -29134       -9654

>exemple_3 [11 35]
ACACGATGCGAGCAATCAAATTTCA
                                  seq   start  end    dG max       dG 3'
5' ACACGATGCGAGCAATCAAATTTCA 3' primer
||||||  |||||||||||| ||
3' TGTGCTTCGCTCGTTAGTTGAACTG 5' genome 54     78     -23644       -2054

>exemple_3 [13 37]
ACGATGCGAGCAATCAAATTTCAGG
                                  seq   start  end    dG max       dG 3'
5' ACGATGCGAGCAATCAAATTTCAGG 3' primer
||||  |||||||||||| ||
3' TGCTTCGCTCGTTAGTTGAACTGGA 5' genome 56     80     -20254       236

>exemple_3 [44 68]
GGATAGAAATACACGAAGCGAGCAA
                                  seq   start  end    dG max       dG 3'
5' GGATAGAAATACACGAAGCGAGCAA 3' primer
|||||||||||||||  ||||||||
3' TCTATCTTTATGTGCTACGCTCGTT 5' genome 1      25     -26614       -10734

>exemple_3 [49 73]
GAAATACACGAAGCGAGCAATCAAC
                                  seq   start  end    dG max       dG 3'
5' GAAATACACGAAGCGAGCAATCAAC 3' primer
||||||||||| ||||||||||||
3' CTTTATGTGCTACGCTCGTTAGTTT 5' genome 6      30     -27174       -8084

>exemple_3 [70 94]
ACTTGACCTAGGTGAGGGATAGGAC
                                  seq   start  end    dG max       dG 3'
5' ACTTGACCTAGGTGAGGGATAGGAC 3' primer
|||||||||||||||||||||||||
3' TGAACTGGATCCACTCCCTATCCTG 5' genome 72     96     -31714       -9574
```

NB1: if the option `[-n]` had been omitted, the program would have generated an empty prime_ok file because all the primers would have had at least one hybridization site (at their own position).

NB2 : the same command must be carried out on the complementary strand of the genome.

4. Visualization of the primers features: ithos_viz

This utility displays the primer features. For each primer, it gives:

- the GC percent
- the melting temperature: Tm
- the maximal suite of identical nucleotides
- the size of the biggest stem-loop structure
- maximal deltaG for the complementary primer
- maximal deltaG in 3' for the complementary primer
- Stability in 5'
- Stability in 3'

Software Execution

The program is launched by the command line:

```
ithos_viz <primers> <parameters>
```

Results are displayed on monitor screen

Input files

- `<primers>` is a text file (FASTA format) that contains a primers list
- `<parameters>` is a text file that contains parameters of the different filters

Example 4: visualization of primer_ok file

The execution of the following command line:

```
ithos_viz primer_ok parameters
```

Displays on the monitor screen:

```
>exemple_3 [59 83]            %GC     Tm      #rep    Hpin    Cp_dG   Cp_dG3' Sta_5' Sta_3'
AAGCGAGCAATCAACTTGACCTAGG     48      58      2       3       -210    1960    -6690   -4980

>exemple_3 [64 88]            %GC     Tm      #rep    Hpin    Cp_dG   Cp_dG3' Sta_5' Sta_3'
AGCAATCAACTTGACCTAGGTGAGG     48      58      2       3       1960    1960    -5970   -5870

>exemple_3 [76 100]           %GC     Tm      #rep    Hpin    Cp_dG   Cp_dG3' Sta_5' Sta_3'
GACCTAGGTGAGGGATAGGACCAGA     56      59      3       4       1960    1960    -5860   -5870
```

5. Parameters file

This is a text file that enables modifying the default parameters of the filters. The same file is read by the three programs.

Example

```
# IThOS PARAMETERS

# primers length

lengthMin 25
lengthMax 25

# filter 1: GC percentage

pcGCMin 40
pcGCMax 60

# filter 2: tm

oligoTmMin 57
oligoTmMax 67
dnaConc 500
saltConc 50

# filter 3: hairpin

maxHpDup 4
maxHpLoop 4

# filter 4: repeat

nbRepeat 6

# filter 5: auto complementarity

maxDeltaGAuto -10000
maxDeltaGAuto3 -6000
sizeDeltaGAuto 8
sizeDeltaGAuto3 8

# filter 6: internal stability to 3' & 5' extremities

sizeExt5 5
sizeExt3 5
deltaG5 -4000
deltaG3min -6000
deltaG3max -3000

#site hybridation (only used by ithos_chk)

sizeDeltaGHybrid3 8
maxDeltaGHybrid3 -12000
maxDeltaGHybrid -18000
```